# Modelling the Effect of Context on Sentence Acceptability

Shalom Lappin*

University of Gothenburg, King's College London, and

Queen Mary University of London

*Joint work with Jey Han Lau (The University of Melbourne), Carlos Santos Armendariz (Queen Mary), Matthew Purver (Queen Mary), and Chang Shu (University of Nottingham Ningbo China)

CLASP Seminar
University of Gothenburg

December 4, 2019

# Outline

# Human Acceptability Judgments

- Lau, Clark, and Lappin (2016) (LCL) use round-trip machine translation (MT) to introduce infelicities into English sentences from the BNC and English Wikipedia sentences.
- They translate through Spanish, Norwegian, Chinese, and Japanese using Google translate.
- They use Amazon Mechanical Turk (AMT) to obtain human acceptability ratings.
- They construct HITS of five randomly selected sentences, with an original non-translated sentence included in each HIT.
- Annotators are filtered for English fluency and reliability.

# Human Acceptability Judgments

- Lau, Clark, and Lappin (2016) (LCL) use round-trip machine translation (MT) to introduce infelicities into English sentences from the BNC and English Wikipedia sentences.
- They translate through Spanish, Norwegian, Chinese, and Japanese using Google translate.
- They use Amazon Mechanical Turk (AMT) to obtain human acceptability ratings.
- They construct HITS of five randomly selected sentences, with an original non-translated sentence included in each HIT.
- Annotators are filtered for English fluency and reliability.

# Human Acceptability Judgments

- Lau, Clark, and Lappin (2016) (LCL) use round-trip machine translation (MT) to introduce infelicities into English sentences from the BNC and English Wikipedia sentences.
- They translate through Spanish, Norwegian, Chinese, and Japanese using Google translate.
- They use Amazon Mechanical Turk (AMT) to obtain human acceptability ratings.
- They construct HITS of five randomly selected sentences, with an original non-translated sentence included in each HIT.
- Annotators are filtered for English fluency and reliability.

# Human Acceptability Judgments

- Lau, Clark, and Lappin (2016) (LCL) use round-trip machine translation (MT) to introduce infelicities into English sentences from the BNC and English Wikipedia sentences.
- They translate through Spanish, Norwegian, Chinese, and Japanese using Google translate.
- They use Amazon Mechanical Turk (AMT) to obtain human acceptability ratings.
- They construct HITS of five randomly selected sentences, with an original non-translated sentence included in each HIT.
- Annotators are filtered for English fluency and reliability.

# Human Acceptability Judgments

- Lau, Clark, and Lappin (2016) (LCL) use round-trip machine translation (MT) to introduce infelicities into English sentences from the BNC and English Wikipedia sentences.
- They translate through Spanish, Norwegian, Chinese, and Japanese using Google translate.
- They use Amazon Mechanical Turk (AMT) to obtain human acceptability ratings.
- They construct HITS of five randomly selected sentences, with an original non-translated sentence included in each HIT.
- Annotators are filtered for English fluency and reliability.

# Gradience in Acceptability Judgements

- LCL obtained 2500 annotated English sentences of 8-25 words for each corpus.
- They used three distinct modes of presentation: binary classification, a four categories of naturalness presentation, and a sliding scale (underlying 100 points).
- There is a high Pearson pairwise correlation of $\geq 0.92$ among the results for these presentations.
- Both aggregate and individual judgments exhibit a significant degree of gradience in acceptability across the annotated sets.

# Gradience in Acceptability Judgements

- LCL obtained 2500 annotated English sentences of 8-25 words for each corpus.
- They used three distinct modes of presentation: binary classification, a four categories of naturalness presentation, and a sliding scale (underlying 100 points).
- There is a high Pearson pairwise correlation of $\geq 0.92$ among the results for these presentations.
- Both aggregate and individual judgments exhibit a significant degree of gradience in acceptability across the annotated sets.

# Gradience in Acceptability Judgements

- LCL obtained 2500 annotated English sentences of 8-25 words for each corpus.
- They used three distinct modes of presentation: binary classification, a four categories of naturalness presentation, and a sliding scale (underlying 100 points).
- There is a high Pearson pairwise correlation of $\geq 0.92$ among the results for these presentations.
- Both aggregate and individual judgments exhibit a significant degree of gradience in acceptability across the annotated sets.

# Gradience in Acceptability Judgements

- LCL obtained 2500 annotated English sentences of 8-25 words for each corpus.
- They used three distinct modes of presentation: binary classification, a four categories of naturalness presentation, and a sliding scale (underlying 100 points).
- There is a high Pearson pairwise correlation of $\geq 0.92$ among the results for these presentations.
- Both aggregate and individual judgments exhibit a significant degree of gradience in acceptability across the annotated sets.

# Predicting Acceptability Judgements

- LCL experiment with a variety of machine learning language models to predict the mean human acceptability judgments of their annotated test sets.
- These include lexical N-grams, a Bayesian Hidden Markov Model (BHMM), a topic driven HMM, a two-tier HMM, and a simple Recurrent Neural Network (RNN).
- They train their models on corpora of 100m words of Wikipedia text in English, German, Spanish, and Russian, respectively.
- LCL use these models to generate a logprob distribution for crowd source annotated test sets in these languages.
- They normalise the logprob values with scoring functions that neutralise the effect of sentence length and word frequency.

# Predicting Acceptability Judgements

- LCL experiment with a variety of machine learning language models to predict the mean human acceptability judgments of their annotated test sets.
- These include lexical N-grams, a Bayesian Hidden Markov Model (BHMM), a topic driven HMM, a two-tier HMM, and a simple Recurrent Neural Network (RNN).
- They train their models on corpora of 100m words of Wikipedia text in English, German, Spanish, and Russian, respectively.
- LCL use these models to generate a logprob distribution for crowd source annotated test sets in these languages.
- They normalise the logprob values with scoring functions that neutralise the effect of sentence length and word frequency.

# Predicting Acceptability Judgements

- LCL experiment with a variety of machine learning language models to predict the mean human acceptability judgments of their annotated test sets.
- These include lexical N-grams, a Bayesian Hidden Markov Model (BHMM), a topic driven HMM, a two-tier HMM, and a simple Recurrent Neural Network (RNN).
- They train their models on corpora of 100m words of Wikipedia text in English, German, Spanish, and Russian, respectively.
- LCL use these models to generate a logprob distribution for crowd source annotated test sets in these languages.
- They normalise the logprob values with scoring functions that neutralise the effect of sentence length and word frequency.

# Predicting Acceptability Judgements

- LCL experiment with a variety of machine learning language models to predict the mean human acceptability judgments of their annotated test sets.
- These include lexical N-grams, a Bayesian Hidden Markov Model (BHMM), a topic driven HMM, a two-tier HMM, and a simple Recurrent Neural Network (RNN).
- They train their models on corpora of 100m words of Wikipedia text in English, German, Spanish, and Russian, respectively.
- LCL use these models to generate a logprob distribution for crowd source annotated test sets in these languages.
- They normalise the logprob values with scoring functions that neutralise the effect of sentence length and word frequency.

# Predicting Acceptability Judgements

- LCL experiment with a variety of machine learning language models to predict the mean human acceptability judgments of their annotated test sets.
- These include lexical N-grams, a Bayesian Hidden Markov Model (BHMM), a topic driven HMM, a two-tier HMM, and a simple Recurrent Neural Network (RNN).
- They train their models on corpora of 100m words of Wikipedia text in English, German, Spanish, and Russian, respectively.
- LCL use these models to generate a logprob distribution for crowd source annotated test sets in these languages.
- They normalise the logprob values with scoring functions that neutralise the effect of sentence length and word frequency.

# Sentence Acceptability Measures

| Scoring Function | Equation |
|---|---|
| *LogProb* = | $\log P_m(\xi)$ |
| *Mean LP* = | $\dfrac{\log P_m(\xi)}{\lvert\xi\rvert}$ |
| *Norm LP (Div)* = | $-\dfrac{\log P_m(\xi)}{\log P_u(\xi)}$ |
| *SLOR* = | $\dfrac{\log P_m(\xi) - \log P_u(\xi)}{\lvert\xi\rvert}$ |

$\xi =$ sentence;
$P_m(\xi) =$ the probability of the sentence given by the model;
$P_u(\xi) =$ is the unigram probability of the sentence;
*SLOR* is proposed by Pauls and Klein (2012)

# Results of the LCL Modelling Experiments

- LCL use the Pearson coefficient to measure the correlation between mean human judgments and a model's prediction of acceptability scores for a test set.
- In general SLOR was the most robustly successful acceptability measure across different test sets.
- The RNN outperformed the other models for all Wikipedia test sets.
- For the English Wikipedia test set it achieved a Pearson coefficient of 0.57 with SLOR, and 0.6 or higher for the other language test sets.

# Results of the LCL Modelling Experiments

- LCL use the Pearson coefficient to measure the correlation between mean human judgments and a model's prediction of acceptability scores for a test set.
- In general SLOR was the most robustly successful acceptability measure across different test sets.
- The RNN outperformed the other models for all Wikipedia test sets.
- For the English Wikipedia test set it achieved a Pearson coefficient of 0.57 with SLOR, and 0.6 or higher for the other language test sets.

# Results of the LCL Modelling Experiments

- LCL use the Pearson coefficient to measure the correlation between mean human judgments and a model's prediction of acceptability scores for a test set.
- In general SLOR was the most robustly successful acceptability measure across different test sets.
- The RNN outperformed the other models for all Wikipedia test sets.
- For the English Wikipedia test set it achieved a Pearson coefficient of 0.57 with SLOR, and 0.6 or higher for the other language test sets.

# Results of the LCL Modelling Experiments

- LCL use the Pearson coefficient to measure the correlation between mean human judgments and a model's prediction of acceptability scores for a test set.
- In general SLOR was the most robustly successful acceptability measure across different test sets.
- The RNN outperformed the other models for all Wikipedia test sets.
- For the English Wikipedia test set it achieved a Pearson coefficient of 0.57 with SLOR, and 0.6 or higher for the other language test sets.

# Modeling Acceptability Independently of Context

- LCL test speakers' acceptability judgments for sentences presented outside of any context beyond the HIT in which they appear.
- The sentences in each HIT are randomly selected.
- LCL's models predict acceptability without reference to context.
- Document context is not explicitly represented in any of the models in either training or testing.

# Modeling Acceptability Independently of Context

- LCL test speakers' acceptability judgments for sentences presented outside of any context beyond the HIT in which they appear.
- The sentences in each HIT are randomly selected.
- LCL's models predict acceptability without reference to context.
- Document context is not explicitly represented in any of the models in either training or testing.

# Modeling Acceptability Independently of Context

- LCL test speakers' acceptability judgments for sentences presented outside of any context beyond the HIT in which they appear.
- The sentences in each HIT are randomly selected.
- LCL's models predict acceptability without reference to context.
- Document context is not explicitly represented in any of the models in either training or testing.

# Modeling Acceptability Independently of Context

- LCL test speakers' acceptability judgments for sentences presented outside of any context beyond the HIT in which they appear.
- The sentences in each HIT are randomly selected.
- LCL's models predict acceptability without reference to context.
- Document context is not explicitly represented in any of the models in either training or testing.

# Human Acceptability Judgments in Context

- Bernardy, Lappin, and Lau (2018) (BLL) constructed two datasets of sentences annotated with acceptability ratings, one judged with and the other without document context.
- They extracted 100 random articles from the English Wikipedia and sampled a sentence from each article.
- They tried LCL's method of using Google Translate (GT) for round-trip MT to generate a set of sentences with varying degrees of acceptability.
- GT has improved to the point that a pilot study indicated that human annotators rated most round-trip translated sentences as highly as English originals.

# Human Acceptability Judgments in Context

- Bernardy, Lappin, and Lau (2018) (BLL) constructed two datasets of sentences annotated with acceptability ratings, one judged with and the other without document context.
- They extracted 100 random articles from the English Wikipedia and sampled a sentence from each article.
- They tried LCL's method of using Google Translate (GT) for round-trip MT to generate a set of sentences with varying degrees of acceptability.
- GT has improved to the point that a pilot study indicated that human annotators rated most round-trip translated sentences as highly as English originals.

# Human Acceptability Judgments in Context

- Bernardy, Lappin, and Lau (2018) (BLL) constructed two datasets of sentences annotated with acceptability ratings, one judged with and the other without document context.
- They extracted 100 random articles from the English Wikipedia and sampled a sentence from each article.
- They tried LCL's method of using Google Translate (GT) for round-trip MT to generate a set of sentences with varying degrees of acceptability.
- GT has improved to the point that a pilot study indicated that human annotators rated most round-trip translated sentences as highly as English originals.

# Human Acceptability Judgments in Context

- Bernardy, Lappin, and Lau (2018) (BLL) constructed two datasets of sentences annotated with acceptability ratings, one judged with and the other without document context.
- They extracted 100 random articles from the English Wikipedia and sampled a sentence from each article.
- They tried LCL's method of using Google Translate (GT) for round-trip MT to generate a set of sentences with varying degrees of acceptability.
- GT has improved to the point that a pilot study indicated that human annotators rated most round-trip translated sentences as highly as English originals.

# The Out-of-Context Annotated Test Set

- As an alternative, BLL used the more traditional statistical phrase based Moses MT system (Koehn et al., 2007).
- They applied the pretrained Moses models for round-trip MT into Czech, Spanish, German and French, and then back to English.
- This provided a distribution of acceptability judgments over sentences comparable to those which LCL obtained in their experiments.
- Following LCL's protocol, BLL used HITS with a four category acceptability rating for crowd source annotation of both their data sets.

# The Out-of-Context Annotated Test Set

- As an alternative, BLL used the more traditional statistical phrase based Moses MT system (Koehn et al., 2007).
- They applied the pretrained Moses models for round-trip MT into Czech, Spanish, German and French, and then back to English.
- This provided a distribution of acceptability judgments over sentences comparable to those which LCL obtained in their experiments.
- Following LCL's protocol, BLL used HITS with a four category acceptability rating for crowd source annotation of both their data sets.

# The Out-of-Context Annotated Test Set

- As an alternative, BLL used the more traditional statistical phrase based Moses MT system (Koehn et al., 2007).
- They applied the pretrained Moses models for round-trip MT into Czech, Spanish, German and French, and then back to English.
- This provided a distribution of acceptability judgments over sentences comparable to those which LCL obtained in their experiments.
- Following LCL's protocol, BLL used HITS with a four category acceptability rating for crowd source annotation of both their data sets.

# The Out-of-Context Annotated Test Set

- As an alternative, BLL used the more traditional statistical phrase based Moses MT system (Koehn et al., 2007).
- They applied the pretrained Moses models for round-trip MT into Czech, Spanish, German and French, and then back to English.
- This provided a distribution of acceptability judgments over sentences comparable to those which LCL obtained in their experiments.
- Following LCL's protocol, BLL used HITS with a four category acceptability rating for crowd source annotation of both their data sets.

# The In-Context Annotated Test Set

- The target sentence was highlighted in boldface, with one preceding and one succeeding sentence included as additional context.
- Annotators had the option of revealing the full document context by clicking on the preceding and succeeding sentences.
- As in the out-of-context test set, sentences were presented in HITS of five, one from the original English set, and four from the round-trip translations.
- Each HIT contained one sentence per target language, with no sentence type appearing more than once in a HIT.

# The In-Context Annotated Test Set

- The target sentence was highlighted in boldface, with one preceding and one succeeding sentence included as additional context.
- Annotators had the option of revealing the full document context by clicking on the preceding and succeeding sentences.
- As in the out-of-context test set, sentences were presented in HITS of five, one from the original English set, and four from the round-trip translations.
- Each HIT contained one sentence per target language, with no sentence type appearing more than once in a HIT.

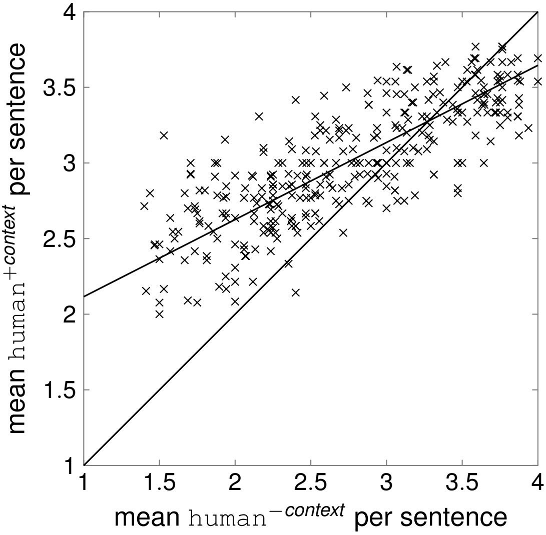
# The In-Context Annotated Test Set

- The target sentence was highlighted in boldface, with one preceding and one succeeding sentence included as additional context.
- Annotators had the option of revealing the full document context by clicking on the preceding and succeeding sentences.
- As in the out-of-context test set, sentences were presented in HITS of five, one from the original English set, and four from the round-trip translations.
- Each HIT contained one sentence per target language, with no sentence type appearing more than once in a HIT.

# The In-Context Annotated Test Set

- The target sentence was highlighted in boldface, with one preceding and one succeeding sentence included as additional context.
- Annotators had the option of revealing the full document context by clicking on the preceding and succeeding sentences.
- As in the out-of-context test set, sentences were presented in HITS of five, one from the original English set, and four from the round-trip translations.
- Each HIT contained one sentence per target language, with no sentence type appearing more than once in a HIT.

# Annotation Results



# Analysing the Effect of Context on Acceptability Judgments

- BLL found a strong Pearson's $r$ correlation of 0.80 between mean out-of-context and in-context judgments.
- The average difference between $\texttt{human}^{-context}$ and $\texttt{human}^{+context}$ is represented by the distance between the linear regression and the full diagonal in the graph.
- These lines cross at $\texttt{human}^{+context} = \texttt{human}^{-context} = 3.28$, the point where context no longer boosts acceptability.

# Analysing the Effect of Context on Acceptability Judgments

- BLL found a strong Pearson's *r* correlation of 0.80 between mean out-of-context and in-context judgments.
- The average difference between $\texttt{human}^{-context}$ and $\texttt{human}^{+context}$ is represented by the distance between the linear regression and the full diagonal in the graph.
- These lines cross at $\texttt{human}^{+context} = \texttt{human}^{-context} = 3.28$, the point where context no longer boosts acceptability.

# Analysing the Effect of Context on Acceptability Judgments

- BLL found a strong Pearson's *r* correlation of 0.80 between mean out-of-context and in-context judgments.
- The average difference between $\texttt{human}^{-context}$ and $\texttt{human}^{+context}$ is represented by the distance between the linear regression and the full diagonal in the graph.
- These lines cross at $\texttt{human}^{+context} = \texttt{human}^{-context} = 3.28$, the point where context no longer boosts acceptability.

# The Compression Effect

- Adding context generally improves acceptability, but the pattern reverses as acceptability approaches maximal mean rating values.
- This “compressess” the distribution of (mean) ratings, pushing the extremes to the middle.
- The net effect of this compression lowers correlation, as the good and bad sentences for the in-context test set are not as clearly separable as they are in the out-of context test set.

# The Compression Effect

- Adding context generally improves acceptability, but the pattern reverses as acceptability approaches maximal mean rating values.
- This “compressess” the distribution of (mean) ratings, pushing the extremes to the middle.
- The net effect of this compression lowers correlation, as the good and bad sentences for the in-context test set are not as clearly separable as they are in the out-of context test set.

# A Similar Pattern for a Different Task

- Bizzoni and Lappin (2019) (BL) test the effect of context on gradient judgments of paraphrase for a metaphorical sentence.
- They solicited AMT crowd source ratings for pairs containing a metaphorical sentence and a candidate for a literal paraphrase of that sentence.
- In one test set 200 pairs are rated on a four category scale of paraphrase appropriateness, independently of context.
- In the second test set the same pairs are judged within a context of a preceding and a following sentence.
- BL observed the same compression effect with the in-context paraphrase judgments that BLL obtained for in-context acceptability ratings.

# A Similar Pattern for a Different Task

- Bizzoni and Lappin (2019) (BL) test the effect of context on gradient judgments of paraphrase for a metaphorical sentence.
- They solicited AMT crowd source ratings for pairs containing a metaphorical sentence and a candidate for a literal paraphrase of that sentence.
- In one test set 200 pairs are rated on a four category scale of paraphrase appropriateness, independently of context.
- In the second test set the same pairs are judged within a context of a preceding and a following sentence.
- BL observed the same compression effect with the in-context paraphrase judgments that BLL obtained for in-context acceptability ratings.

# A Similar Pattern for a Different Task

- Bizzoni and Lappin (2019) (BL) test the effect of context on gradient judgments of paraphrase for a metaphorical sentence.
- They solicited AMT crowd source ratings for pairs containing a metaphorical sentence and a candidate for a literal paraphrase of that sentence.
- In one test set 200 pairs are rated on a four category scale of paraphrase appropriateness, independently of context.
- In the second test set the same pairs are judged within a context of a preceding and a following sentence.
- BL observed the same compression effect with the in-context paraphrase judgments that BLL obtained for in-context acceptability ratings.

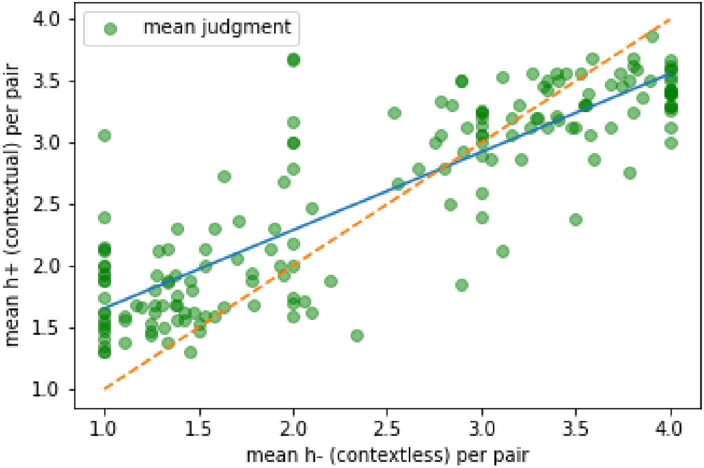
# A Similar Pattern for a Different Task

- Bizzoni and Lappin (2019) (BL) test the effect of context on gradient judgments of paraphrase for a metaphorical sentence.
- They solicited AMT crowd source ratings for pairs containing a metaphorical sentence and a candidate for a literal paraphrase of that sentence.
- In one test set 200 pairs are rated on a four category scale of paraphrase appropriateness, independently of context.
- In the second test set the same pairs are judged within a context of a preceding and a following sentence.
- BL observed the same compression effect with the in-context paraphrase judgments that BLL obtained for in-context acceptability ratings.

# A Similar Pattern for a Different Task

- Bizzoni and Lappin (2019) (BL) test the effect of context on gradient judgments of paraphrase for a metaphorical sentence.
- They solicited AMT crowd source ratings for pairs containing a metaphorical sentence and a candidate for a literal paraphrase of that sentence.
- In one test set 200 pairs are rated on a four category scale of paraphrase appropriateness, independently of context.
- In the second test set the same pairs are judged within a context of a preceding and a following sentence.
- BL observed the same compression effect with the in-context paraphrase judgments that BLL obtained for in-context acceptability ratings.

# BL's Regression Graph for Paraphrase Judgments



# Two DNN LMs

- BLL experiment with two Deep Neural Network Language models to predict the human sentence ratings for each of their test sets.
- `lstm` (Hochreiter and Schmidhuber (1997)), Mikolov et al. (2010)) is a standard LSTM language model, trained over a corpus to predict word sequences.
- `tdlm` (Lau et al. (2017)) is a topic driven neural LM.
- The topic model component of `tdlm` produces topics by processing documents through a convolutional layer and aligning it with trainable topic embeddings.
- The language model component of `tdlm` incorporates context by combining its topic vector with its LSTM's hidden state, to generate the probability distribution for the next word.

# Two DNN LMs

- BLL experiment with two Deep Neural Network Language models to predict the human sentence ratings for each of their test sets.
- `lstm` (Hochreiter and Schmidhuber (1997)), Mikolov et al. (2010)) is a standard LSTM language model, trained over a corpus to predict word sequences.
- `tdlm` (Lau et al. (2017)) is a topic driven neural LM.
- The topic model component of `tdlm` produces topics by processing documents through a convolutional layer and aligning it with trainable topic embeddings.
- The language model component of `tdlm` incorporates context by combining its topic vector with its LSTM's hidden state, to generate the probability distribution for the next word.

# Two DNN LMs

- BLL experiment with two Deep Neural Network Language models to predict the human sentence ratings for each of their test sets.
- `lstm` (Hochreiter and Schmidhuber (1997)), Mikolov et al. (2010)) is a standard LSTM language model, trained over a corpus to predict word sequences.
- `tdlm` (Lau et al. (2017)) is a topic driven neural LM.
- The topic model component of `tdlm` produces topics by processing documents through a convolutional layer and aligning it with trainable topic embeddings.
- The language model component of `tdlm` incorporates context by combining its topic vector with its LSTM's hidden state, to generate the probability distribution for the next word.

# Two DNN LMs

- BLL experiment with two Deep Neural Network Language models to predict the human sentence ratings for each of their test sets.
- `lstm` (Hochreiter and Schmidhuber (1997)), Mikolov et al. (2010)) is a standard LSTM language model, trained over a corpus to predict word sequences.
- `tdlm` (Lau et al. (2017)) is a topic driven neural LM.
- The topic model component of `tdlm` produces topics by processing documents through a convolutional layer and aligning it with trainable topic embeddings.
- The language model component of `tdlm` incorporates context by combining its topic vector with its LSTM's hidden state, to generate the probability distribution for the next word.

# Two DNN LMs

- BLL experiment with two Deep Neural Network Language models to predict the human sentence ratings for each of their test sets.
- `lstm` (Hochreiter and Schmidhuber (1997)), Mikolov et al. (2010)) is a standard LSTM language model, trained over a corpus to predict word sequences.
- `tdlm` (Lau et al. (2017)) is a topic driven neural LM.
- The topic model component of `tdlm` produces topics by processing documents through a convolutional layer and aligning it with trainable topic embeddings.
- The language model component of `tdlm` incorporates context by combining its topic vector with its LSTM's hidden state, to generate the probability distribution for the next word.

# Four LM Variants

- Both LMs can use the document context as a prefix input to the sentence at test time.
- This gives us 4 variant LMs at test time.
  1. `lstm`$^{-c}$ and `tdlm`$^{-c}$, which use only sentences from a test set as input.
  2. `lstm`$^{+c}$ and `tdlm`$^{+c}$, which use sentence and context at test time.
- To map sentence probability to acceptability we use LCL's 3 scoring functions.

# Four LM Variants

- Both LMs can use the document context as a prefix input to the sentence at test time.
- This gives us 4 variant LMs at test time.
  1. $\texttt{lstm}^{-c}$ and $\texttt{tdlm}^{-c}$, which use only sentences from a test set as input.
  2. $\texttt{lstm}^{+c}$ and $\texttt{tdlm}^{+c}$, which use sentence and context at test time.
- To map sentence probability to acceptability we use LCL's 3 scoring functions.

# Four LM Variants

- Both LMs can use the document context as a prefix input to the sentence at test time.
- This gives us 4 variant LMs at test time.
  1. $\texttt{lstm}^{-c}$ and $\texttt{tdlm}^{-c}$, which use only sentences from a test set as input.
  2. $\texttt{lstm}^{+c}$ and $\texttt{tdlm}^{+c}$, which use sentence and context at test time.
- To map sentence probability to acceptability we use LCL's 3 scoring functions.

# Training and Evaluation of the LMs

- BLL train `tdlm` and `lstm` on a sample of 100K English Wikipedia articles, which has no overlap with the 100 documents used for test set annotation.
- The training data has approximately 40M tokens and a vocabulary size of 66K
- To assess the performance of the acceptability measures, BLL compute Pearson's *r* against mean human ratings.
- BLL also experimented with Spearman's rank correlation, but found similar trends, and so they present only the Pearson results.

# Training and Evaluation of the LMs

- BLL train `tdlm` and `lstm` on a sample of 100K English Wikipedia articles, which has no overlap with the 100 documents used for test set annotation.
- The training data has approximately 40M tokens and a vocabulary size of 66K
- To assess the performance of the acceptability measures, BLL compute Pearson's *r* against mean human ratings.
- BLL also experimented with Spearman's rank correlation, but found similar trends, and so they present only the Pearson results.

# Training and Evaluation of the LMs

- BLL train `tdlm` and `lstm` on a sample of 100K English Wikipedia articles, which has no overlap with the 100 documents used for test set annotation.
- The training data has approximately 40M tokens and a vocabulary size of 66K
- To assess the performance of the acceptability measures, BLL compute Pearson's *r* against mean human ratings.
- BLL also experimented with Spearman's rank correlation, but found similar trends, and so they present only the Pearson results.

# Training and Evaluation of the LMs

- BLL train `tdlm` and `lstm` on a sample of 100K English Wikipedia articles, which has no overlap with the 100 documents used for test set annotation.
- The training data has approximately 40M tokens and a vocabulary size of 66K
- To assess the performance of the acceptability measures, BLL compute Pearson's *r* against mean human ratings.
- BLL also experimented with Spearman's rank correlation, but found similar trends, and so they present only the Pearson results.

# Model Performance on the Prediction Task

| Rtg | Model | *LP* | *Mean* | *NrmD* | *SLOR* |
|---|---|---|---|---|---|
| $\texttt{human}^{-context}$ | $\texttt{lstm}^{-c}$ | 0.151 | 0.487 | **0.586** | 0.584 |
| | $\texttt{lstm}^{+c}$ | 0.161 | 0.529 | 0.618 | **0.633** |
| | $\texttt{tdlm}^{-c}$ | 0.147 | 0.515 | 0.634 | **0.640** |
| | $\texttt{tdlm}^{+c}$ | 0.165 | 0.541 | 0.645 | **0.653** |
| $\texttt{human}^{+context}$ | $\texttt{lstm}^{-c}$ | 0.153 | 0.421 | 0.494 | **0.503** |
| | $\texttt{lstm}^{+c}$ | 0.168 | 0.459 | 0.522 | **0.546** |
| | $\texttt{tdlm}^{-c}$ | 0.153 | 0.450 | 0.541 | **0.557** |
| | $\texttt{tdlm}^{+c}$ | 0.169 | 0.473 | 0.552 | **0.568** |

# Discussion of the Models' Performance

- `lstm`$^{-c}$ against `human`$^{-context}$ with SLOR achieves 0.584, slightly surpassing the performance of the RNN with SLOR in the original LCL experiment (0.570).
- Across all models (`lstm` and `tdlm`) and human ratings (`human`$^{-context}$ and `human`$^{+context}$), using context at test time improves model performance.
- Taking context into account helps in modelling acceptability, regardless of whether it is tested against judgements made with (`human`$^{+context}$) or without context (`human`$^{-context}$).
- `tdlm` consistently outperforms `lstm` over both types of human ratings and test input variants.
- Context helps in the modelling of acceptability, whether it is incorporated during training (`lstm` vs. `tdlm`) or at test time (`lstm`$^{-c}$/`tdlm`$^{-c}$ vs. `lstm`$^{+c}$/`tdlm`$^{+c}$).

# Discussion of the Models' Performance

- `lstm`$^{-c}$ against `human`$^{-context}$ with SLOR achieves 0.584, slightly surpassing the performance of the RNN with SLOR in the original LCL experiment (0.570).
- Across all models (`lstm` and `tdlm`) and human ratings (`human`$^{-context}$ and `human`$^{+context}$), using context at test time improves model performance.
- Taking context into account helps in modelling acceptability, regardless of whether it is tested against judgements made with (`human`$^{+context}$) or without context (`human`$^{-context}$).
- `tdlm` consistently outperforms `lstm` over both types of human ratings and test input variants.
- Context helps in the modelling of acceptability, whether it is incorporated during training (`lstm` vs. `tdlm`) or at test time (`lstm`$^{-c}$/`tdlm`$^{-c}$ vs. `lstm`$^{+c}$/`tdlm`$^{+c}$).

# Discussion of the Models' Performance

- `lstm`$^{-c}$ against `human`$^{-context}$ with SLOR achieves 0.584, slightly surpassing the performance of the RNN with SLOR in the original LCL experiment (0.570).
- Across all models (`lstm` and `tdlm`) and human ratings (`human`$^{-context}$ and `human`$^{+context}$), using context at test time improves model performance.
- Taking context into account helps in modelling acceptability, regardless of whether it is tested against judgements made with (`human`$^{+context}$) or without context (`human`$^{-context}$).
- `tdlm` consistently outperforms `lstm` over both types of human ratings and test input variants.
- Context helps in the modelling of acceptability, whether it is incorporated during training (`lstm` vs. `tdlm`) or at test time (`lstm`$^{-c}$/`tdlm`$^{-c}$ vs. `lstm`$^{+c}$/`tdlm`$^{+c}$).

# Discussion of the Models' Performance

- $\texttt{lstm}^{-c}$ against $\texttt{human}^{-context}$ with SLOR achieves 0.584, slightly surpassing the performance of the RNN with SLOR in the original LCL experiment (0.570).
- Across all models (`lstm` and `tdlm`) and human ratings ($\texttt{human}^{-context}$ and $\texttt{human}^{+context}$), using context at test time improves model performance.
- Taking context into account helps in modelling acceptability, regardless of whether it is tested against judgements made with ($\texttt{human}^{+context}$) or without context ($\texttt{human}^{-context}$).
- `tdlm` consistently outperforms `lstm` over both types of human ratings and test input variants.
- Context helps in the modelling of acceptability, whether it is incorporated during training (`lstm` vs. `tdlm`) or at test time ($\texttt{lstm}^{-c}$/$\texttt{tdlm}^{-c}$ vs. $\texttt{lstm}^{+c}$/$\texttt{tdlm}^{+c}$).

# Discussion of the Models' Performance

- lstm$^{-c}$ against human$^{-context}$ with SLOR achieves 0.584, slightly surpassing the performance of the RNN with SLOR in the original LCL experiment (0.570).
- Across all models (lstm and tdlm) and human ratings (human$^{-context}$ and human$^{+context}$), using context at test time improves model performance.
- Taking context into account helps in modelling acceptability, regardless of whether it is tested against judgements made with (human$^{+context}$) or without context (human$^{-context}$).
- tdlm consistently outperforms lstm over both types of human ratings and test input variants.
- Context helps in the modelling of acceptability, whether it is incorporated during training (lstm vs. tdlm) or at test time (lstm$^{-c}$/tdlm$^{-c}$ vs. lstm$^{+c}$/tdlm$^{+c}$).

# The Models' In-Context Predictions

- The *SLOR* correlation of $\texttt{lstm}^{+c}$/$\texttt{tdlm}^{+c}$ vs. $\texttt{human}^{+context}$ (0.546/568) is lower than that of $\texttt{lstm}^{-c}$/$\texttt{tdlm}^{-c}$ vs. $\texttt{human}^{-context}$ (0.584/0.640).

- $\texttt{human}^{+context}$ ratings are more difficult to predict than $\texttt{human}^{-context}$.

# The Models' In-Context Predictions

- The *SLOR* correlation of $\texttt{lstm}^{+c}$/$\texttt{tdlm}^{+c}$ vs. $\texttt{human}^{+context}$ (0.546/568) is lower than that of $\texttt{lstm}^{-c}$/$\texttt{tdlm}^{-c}$ vs. $\texttt{human}^{-context}$ (0.584/0.640).

- $\texttt{human}^{+context}$ ratings are more difficult to predict than $\texttt{human}^{-context}$.

# One Explanation: Discourse Coherence

- But the question remains as to why context reduces the spread between ratings.
- One possible explanation is that annotators focus more on discourse coherence when rating sentences in a document context.
- The issue of discourse coherence does not arise in $\texttt{human}^{-context}$ judgments.
- If this factor is, in fact, significant in annotation, then syntactic infelicities introduced by round-trip MT may play less of a role in rating for the $\texttt{human}^{+context}$ set.

# One Explanation: Discourse Coherence

- But the question remains as to why context reduces the spread between ratings.
- One possible explanation is that annotators focus more on discourse coherence when rating sentences in a document context.
- The issue of discourse coherence does not arise in $\texttt{human}^{-context}$ judgments.
- If this factor is, in fact, significant in annotation, then syntactic infelicities introduced by round-trip MT may play less of a role in rating for the $\texttt{human}^{+context}$ set.

# One Explanation: Discourse Coherence

- But the question remains as to why context reduces the spread between ratings.
- One possible explanation is that annotators focus more on discourse coherence when rating sentences in a document context.
- The issue of discourse coherence does not arise in $\texttt{human}^{-\textit{context}}$ judgments.
- If this factor is, in fact, significant in annotation, then syntactic infelicities introduced by round-trip MT may play less of a role in rating for the $\texttt{human}^{+\textit{context}}$ set.

# One Explanation: Discourse Coherence

- But the question remains as to why context reduces the spread between ratings.
- One possible explanation is that annotators focus more on discourse coherence when rating sentences in a document context.
- The issue of discourse coherence does not arise in $\texttt{human}^{-context}$ judgments.
- If this factor is, in fact, significant in annotation, then syntactic infelicities introduced by round-trip MT may play less of a role in rating for the $\texttt{human}^{+context}$ set.

# A Second Explanation: General Cognitive Load

- A second explanation is that context imposes additional cognitive load (Sweller, 1988; Ito et al., 2018; Causse et al., 2016; Park et al., 2013), which reduces the speaker/hearer's resources for identifying syntactic and semantic anomaly in an individual sentence.
- If the discourse coherence account is correct, then we would expect the compression effect to be prominent with coherent contexts, but not with random contexts, which prevent integration of the sentence into a discourse unit.
- By contrast, the general cognitive load explanation predicts that the compression effect should be observable for both types of context, as each of them causes distraction through use of additional processing resources.

# A Second Explanation: General Cognitive Load

- A second explanation is that context imposes additional cognitive load (Sweller, 1988; Ito et al., 2018; Causse et al., 2016; Park et al., 2013), which reduces the speaker/hearer's resources for identifying syntactic and semantic anomaly in an individual sentence.
- If the discourse coherence account is correct, then we would expect the compression effect to be prominent with coherent contexts, but not with random contexts, which prevent integration of the sentence into a discourse unit.
- By contrast, the general cognitive load explanation predicts that the compression effect should be observable for both types of context, as each of them causes distraction through use of additional processing resources.

# A Second Explanation: General Cognitive Load

- A second explanation is that context imposes additional cognitive load (Sweller, 1988; Ito et al., 2018; Causse et al., 2016; Park et al., 2013), which reduces the speaker/hearer's resources for identifying syntactic and semantic anomaly in an individual sentence.
- If the discourse coherence account is correct, then we would expect the compression effect to be prominent with coherent contexts, but not with random contexts, which prevent integration of the sentence into a discourse unit.
- By contrast, the general cognitive load explanation predicts that the compression effect should be observable for both types of context, as each of them causes distraction through use of additional processing resources.

# Our Test Set

- Following BLL's protocol, we generated a test set of 250 sentences from 50 English Wikipedia sentences, through round trip MT, with Moses.
- We split the test set into 25 HITs of 10 sentences.
- Each HIT contains 2 original English sentences and 8 translated sentences, which are different from each other and not derived from either of the originals.
- We used AMT crowd sourcing to annotate the sentences for naturalness on a four point scale, for three types of context.

# Our Test Set

- Following BLL's protocol, we generated a test set of 250 sentences from 50 English Wikipedia sentences, through round trip MT, with Moses.
- We split the test set into 25 HITs of 10 sentences.
- Each HIT contains 2 original English sentences and 8 translated sentences, which are different from each other and not derived from either of the originals.
- We used AMT crowd sourcing to annotate the sentences for naturalness on a four point scale, for three types of context.

# Our Test Set

- Following BLL's protocol, we generated a test set of 250 sentences from 50 English Wikipedia sentences, through round trip MT, with Moses.
- We split the test set into 25 HITs of 10 sentences.
- Each HIT contains 2 original English sentences and 8 translated sentences, which are different from each other and not derived from either of the originals.
- We used AMT crowd sourcing to annotate the sentences for naturalness on a four point scale, for three types of context.

# Our Test Set

- Following BLL's protocol, we generated a test set of 250 sentences from 50 English Wikipedia sentences, through round trip MT, with Moses.
- We split the test set into 25 HITs of 10 sentences.
- Each HIT contains 2 original English sentences and 8 translated sentences, which are different from each other and not derived from either of the originals.
- We used AMT crowd sourcing to annotate the sentences for naturalness on a four point scale, for three types of context.

# Null, Real, and Random, Contexts

- We presented the sentences in each HIT in null, real, and, random contexts, respectively.
- Each context experiment was performed by a disjoint group of annotators.
- The real contexts consists of the three sentences that immediately precede a sentence in its document.
- The random contexts are consecutive sequences of three sentences taken from other documents.

# Null, Real, and Random, Contexts

- We presented the sentences in each HIT in null, real, and, random contexts, respectively.
- Each context experiment was performed by a disjoint group of annotators.
- The real contexts consists of the three sentences that immediately precede a sentence in its document.
- The random contexts are consecutive sequences of three sentences taken from other documents.

# Null, Real, and Random, Contexts

- We presented the sentences in each HIT in null, real, and, random contexts, respectively.
- Each context experiment was performed by a disjoint group of annotators.
- The real contexts consists of the three sentences that immediately precede a sentence in its document.
- The random contexts are consecutive sequences of three sentences taken from other documents.

# Null, Real, and Random, Contexts

- We presented the sentences in each HIT in null, real, and, random contexts, respectively.
- Each context experiment was performed by a disjoint group of annotators.
- The real contexts consists of the three sentences that immediately precede a sentence in its document.
- The random contexts are consecutive sequences of three sentences taken from other documents.

# A Topic Identification Task

- In the context experiments we first show the context paragraph, and we ask users to select the most appropriate description of its topic from a list of 4 candidate topics.
- Each candidate topic is represented by three words generated with a topic model.
- After performing this task the annotator is shown the sentence to be rated for acceptability.
- This experimental set up insures that annotators read the context sentences before assessing the sentences of the HIT.

# A Topic Identification Task

- In the context experiments we first show the context paragraph, and we ask users to select the most appropriate description of its topic from a list of 4 candidate topics.
- Each candidate topic is represented by three words generated with a topic model.
- After performing this task the annotator is shown the sentence to be rated for acceptability.
- This experimental set up insures that annotators read the context sentences before assessing the sentences of the HIT.

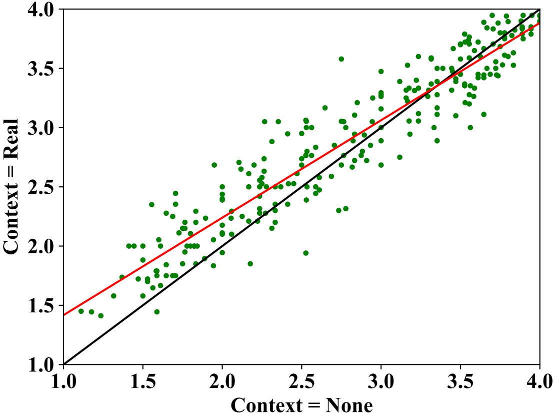
# A Topic Identification Task

- In the context experiments we first show the context paragraph, and we ask users to select the most appropriate description of its topic from a list of 4 candidate topics.
- Each candidate topic is represented by three words generated with a topic model.
- After performing this task the annotator is shown the sentence to be rated for acceptability.
- This experimental set up insures that annotators read the context sentences before assessing the sentences of the HIT.

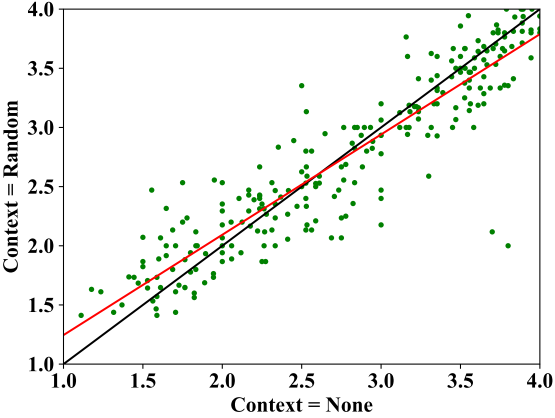
# A Topic Identification Task

- In the context experiments we first show the context paragraph, and we ask users to select the most appropriate description of its topic from a list of 4 candidate topics.
- Each candidate topic is represented by three words generated with a topic model.
- After performing this task the annotator is shown the sentence to be rated for acceptability.
- This experimental set up insures that annotators read the context sentences before assessing the sentences of the HIT.

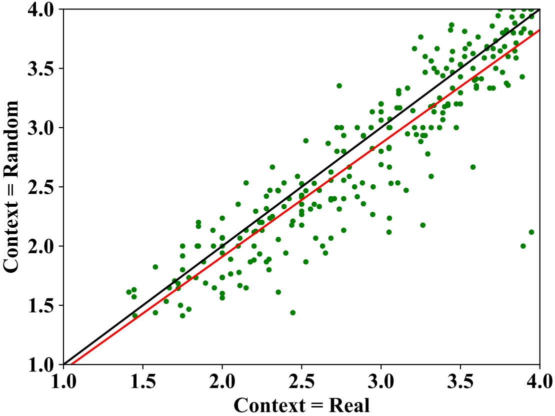
# Human Acceptability Judgments: Real Contexts vs No Contexts



# Human Acceptability Judgments: Random Contexts vs No Contexts



# Human Acceptability Judgments: Random Contexts vs Real Contexts



# Explaining the Compression and Raising Effects

- The compression effect appears in both the $\mathrm{h}^{+}$ (real context) vs. $\mathrm{h}^{\varnothing}$ (null context), and the $\mathrm{h}^{-}$ (random context) vs. $\mathrm{h}^{\varnothing}$ cases.
- In addition, the $\mathrm{h}^{+}$ vs. $\mathrm{h}^{\varnothing}$ regression diagram exhibits a raising effect in real contexts, which pushes the cross over point towards the upper end of the scale.
- In the $\mathrm{h}^{-}$ vs. $\mathrm{h}^{+}$ figure the regression line is parallel to and below the diagonal, indicating a consistent decrease in acceptability ratings from $\mathrm{h}^{+}$ to $\mathrm{h}^{-}$.
- These effects suggest that the cognitive load of processing contexts produces compression in both $\mathrm{h}^{+}$ and $\mathrm{h}^{\varnothing}$, while discourse coherence operates only in $\mathrm{h}^{+}$ to generate a raising of acceptability ratings.

# Explaining the Compression and Raising Effects

- The compression effect appears in both the $\mathrm{h}^{+}$ (real context) vs. $\mathrm{h}^{\varnothing}$ (null context), and the $\mathrm{h}^{-}$ (random context) vs. $\mathrm{h}^{\varnothing}$ cases.
- In addition, the $\mathrm{h}^{+}$ vs. $\mathrm{h}^{\varnothing}$ regression diagram exhibits a raising effect in real contexts, which pushes the cross over point towards the upper end of the scale.
- In the $\mathrm{h}^{-}$ vs. $\mathrm{h}^{+}$ figure the regression line is parallel to and below the diagonal, indicating a consistent decrease in acceptability ratings from $\mathrm{h}^{+}$ to $\mathrm{h}^{-}$.
- These effects suggest that the cognitive load of processing contexts produces compression in both $\mathrm{h}^{+}$ and $\mathrm{h}^{\varnothing}$, while discourse coherence operates only in $\mathrm{h}^{+}$ to generate a raising of acceptability ratings.

# Explaining the Compression and Raising Effects

- The compression effect appears in both the $\mathtt{h}^+$ (real context) vs. $\mathtt{h}^\varnothing$ (null context), and the $\mathtt{h}^-$ (random context) vs. $\mathtt{h}^\varnothing$ cases.
- In addition, the $\mathtt{h}^+$ vs. $\mathtt{h}^\varnothing$ regression diagram exhibits a raising effect in real contexts, which pushes the cross over point towards the upper end of the scale.
- In the $\mathtt{h}^-$ vs. $\mathtt{h}^+$ figure the regression line is parallel to and below the diagonal, indicating a consistent decrease in acceptability ratings from $\mathtt{h}^+$ to $\mathtt{h}^-$.
- These effects suggest that the cognitive load of processing contexts produces compression in both $\mathtt{h}^+$ and $\mathtt{h}^\varnothing$, while discourse coherence operates only in $\mathtt{h}^+$ to generate a raising of acceptability ratings.

# Explaining the Compression and Raising Effects

- The compression effect appears in both the $\mathtt{h}^+$ (real context) vs. $\mathtt{h}^\varnothing$ (null context), and the $\mathtt{h}^-$ (random context) vs. $\mathtt{h}^\varnothing$ cases.
- In addition, the $\mathtt{h}^+$ vs. $\mathtt{h}^\varnothing$ regression diagram exhibits a raising effect in real contexts, which pushes the cross over point towards the upper end of the scale.
- In the $\mathtt{h}^-$ vs. $\mathtt{h}^+$ figure the regression line is parallel to and below the diagonal, indicating a consistent decrease in acceptability ratings from $\mathtt{h}^+$ to $\mathtt{h}^-$.
- These effects suggest that the cognitive load of processing contexts produces compression in both $\mathtt{h}^+$ and $\mathtt{h}^\varnothing$, while discourse coherence operates only in $\mathtt{h}^+$ to generate a raising of acceptability ratings.

# Statistical Significance of the Compression and Discourse Coherence Effects I

- The mean ratings in all three test sets correlate strongly with each other, with Pearson's *r* for $\mathrm{h}^{+}$ vs. $\mathrm{h}^{\varnothing} = 0.945$, $\mathrm{h}^{-}$ vs. $\mathrm{h}^{\varnothing} = 0.917$, and $\mathrm{h}^{-}$ vs. $\mathrm{h}^{+} = 0.901$.
- We used the non-parametric Wilcoxon signed-rank test (one-tailed) to compare the difference between $\mathrm{h}^{+}$ and $\mathrm{h}^{-}$.
- The test gives a *p*-value of $2.4 \times 10^{-8}$, indicating that the discourse coherence effect is significant.

# Statistical Significance of the Compression and Discourse Coherence Effects I

- The mean ratings in all three test sets correlate strongly with each other, with Pearson's *r* for $\mathtt{h}^+$ vs. $\mathtt{h}^\varnothing = 0.945$, $\mathtt{h}^-$ vs. $\mathtt{h}^\varnothing = 0.917$, and $\mathtt{h}^-$ vs. $\mathtt{h}^+ = 0.901$.
- We used the non-parametric Wilcoxon signed-rank test (one-tailed) to compare the difference between $\mathtt{h}^+$ and $\mathtt{h}^-$.
- The test gives a *p*-value of $2.4 \times 10^{-8}$, indicating that the discourse coherence effect is significant.

# Statistical Significance of the Compression and Discourse Coherence Effects I

- The mean ratings in all three test sets correlate strongly with each other, with Pearson's *r* for $\mathtt{h}^{+}$ vs. $\mathtt{h}^{\varnothing} = 0.945$, $\mathtt{h}^{-}$ vs. $\mathtt{h}^{\varnothing} = 0.917$, and $\mathtt{h}^{-}$ vs. $\mathtt{h}^{+} = 0.901$.

- We used the non-parametric Wilcoxon signed-rank test (one-tailed) to compare the difference between $\mathtt{h}^{+}$ and $\mathtt{h}^{-}$.

- The test gives a *p*-value of $2.4 \times 10^{-8}$, indicating that the discourse coherence effect is significant.

# Statistical Significance of the Compression and Discourse Coherence Effects II

- We also used the Wilcoxon test to compare the regression lines for $\mathrm{h}^{+}$ vs. $\mathrm{h}^{\varnothing}$, and $\mathrm{h}^{-}$ vs. $\mathrm{h}^{\varnothing}$, to see if their offsets (constants) and slopes (coefficients) are statistically different.
- The *p*-value for the offset is $2.1 \times 10^{-2}$, confirming that there is a significant discourse coherence effect.
- The *p*-value for the slope, however, is $3.9 \times 10^{-1}$, suggesting that cognitive load compresses the ratings in a consistent way for both $\mathrm{h}^{+}$ and $\mathrm{h}^{-}$, relative to $\mathrm{h}^{\varnothing}$.

# Statistical Significance of the Compression and Discourse Coherence Effects II

- We also used the Wilcoxon test to compare the regression lines for $\mathtt{h}^{+}$ vs. $\mathtt{h}^{\varnothing}$, and $\mathtt{h}^{-}$ vs. $\mathtt{h}^{\varnothing}$, to see if their offsets (constants) and slopes (coefficients) are statistically different.
- The *p*-value for the offset is $2.1 \times 10^{-2}$, confirming that there is a significant discourse coherence effect.
- The *p*-value for the slope, however, is $3.9 \times 10^{-1}$, suggesting that cognitive load compresses the ratings in a consistent way for both $\mathtt{h}^{+}$ and $\mathtt{h}^{-}$, relative to $\mathtt{h}^{\varnothing}$.

# Statistical Significance of the Compression and Discourse Coherence Effects II

- We also used the Wilcoxon test to compare the regression lines for $\mathtt{h}^{+}$ vs. $\mathtt{h}^{\varnothing}$, and $\mathtt{h}^{-}$ vs. $\mathtt{h}^{\varnothing}$, to see if their offsets (constants) and slopes (coefficients) are statistically different.
- The *p*-value for the offset is $2.1 \times 10^{-2}$, confirming that there is a significant discourse coherence effect.
- The *p*-value for the slope, however, is $3.9 \times 10^{-1}$, suggesting that cognitive load compresses the ratings in a consistent way for both $\mathtt{h}^{+}$ and $\mathtt{h}^{-}$, relative to $\mathtt{h}^{\varnothing}$.

# Predicting Sentence Acceptability with Deep Neural Language Models

- In addition to `lstm` and `tdlm` we experiment with three transformer language models (LMs).
- These are `gpt2` (Radford et al., 2019), `bert` (Devlin et al., 2019), and `xlnet` (Yang et al., 2019).
- These models are equipped with large pre-trained lexical embeddings, and they apply multiple self-attention heads to all input words.
- `bert` processes input strings without regard to sequence, in a massively parallel way, which permits it to efficiently identify large numbers of co-occurrence dependency patterns among the words of a string.

# Predicting Sentence Acceptability with Deep Neural Language Models

- In addition to `lstm` and `tdlm` we experiment with three transformer language models (LMs).
- These are `gpt2` (Radford et al., 2019), `bert` (Devlin et al., 2019), and `xlnet` (Yang et al., 2019).
- These models are equipped with large pre-trained lexical embeddings, and they apply multiple self-attention heads to all input words.
- `bert` processes input strings without regard to sequence, in a massively parallel way, which permits it to efficiently identify large numbers of co-occurrence dependency patterns among the words of a string.

# Predicting Sentence Acceptability with Deep Neural Language Models

- In addition to `lstm` and `tdlm` we experiment with three transformer language models (LMs).
- These are `gpt2` (Radford et al., 2019), `bert` (Devlin et al., 2019), and `xlnet` (Yang et al., 2019).
- These models are equipped with large pre-trained lexical embeddings, and they apply multiple self-attention heads to all input words.
- `bert` processes input strings without regard to sequence, in a massively parallel way, which permits it to efficiently identify large numbers of co-occurrence dependency patterns among the words of a string.

# Predicting Sentence Acceptability with Deep Neural Language Models

- In addition to `lstm` and `tdlm` we experiment with three transformer language models (LMs).
- These are `gpt2` (Radford et al., 2019), `bert` (Devlin et al., 2019), and `xlnet` (Yang et al., 2019).
- These models are equipped with large pre-trained lexical embeddings, and they apply multiple self-attention heads to all input words.
- `bert` processes input strings without regard to sequence, in a massively parallel way, which permits it to efficiently identify large numbers of co-occurrence dependency patterns among the words of a string.

# Sentence Probabilities for Transformers

- `lstm` and `gpt2` are unidirectional, and so they can be used to compute the probability of a sentence left to right, according to the formula $\overrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i|w_{<i})$.
- `bert` is bidirectional, and predicts words for both their left and right contexts.
- It requires the formula $\overleftrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i|w_{<i}, w_{>i})$.
- This equation does not yield true probabilities, as its values cannot be normalised to sum to 1.
- Instead these values provide confidence scores of likelihood.
- `xlnet` can be applied either unidirectionally or bidirectionally.

# Sentence Probabilities for Transformers

- `lstm` and `gpt2` are unidirectional, and so they can be used to compute the probability of a sentence left to right, according to the formula $\overrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i})$.
- `bert` is bidirectional, and predicts words for both their left and right contexts.
- It requires the formula $\overleftrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i}, w_{>i})$.
- This equation does not yield true probabilities, as its values cannot be normalised to sum to 1.
- Instead these values provide confidence scores of likelihood.
- `xlnet` can be applied either unidirectionally or bidirectionally.

# Sentence Probabilities for Transformers

- `lstm` and `gpt2` are unidirectional, and so they can be used to compute the probability of a sentence left to right, according to the formula $\overrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i})$.
- `bert` is bidirectional, and predicts words for both their left and right contexts.
- It requires the formula $\overset{\leftrightarrow}{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i}, w_{>i})$.
- This equation does not yield true probabilities, as its values cannot be normalised to sum to 1.
- Instead these values provide confidence scores of likelihood.
- `xlnet` can be applied either unidirectionally or bidirectionally.

# Sentence Probabilities for Transformers

- `lstm` and `gpt2` are unidirectional, and so they can be used to compute the probability of a sentence left to right, according to the formula $\overrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i|w_{<i})$.
- `bert` is bidirectional, and predicts words for both their left and right contexts.
- It requires the formula $\overleftrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i|w_{<i}, w_{>i})$.
- This equation does not yield true probabilities, as its values cannot be normalised to sum to 1.
- Instead these values provide confidence scores of likelihood.
- `xlnet` can be applied either unidirectionally or bidirectionally.

# Sentence Probabilities for Transformers

- `lstm` and `gpt2` are unidirectional, and so they can be used to compute the probability of a sentence left to right, according to the formula $\overrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i})$.
- `bert` is bidirectional, and predicts words for both their left and right contexts.
- It requires the formula $\overleftrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i}, w_{>i})$.
- This equation does not yield true probabilities, as its values cannot be normalised to sum to 1.
- Instead these values provide confidence scores of likelihood.
- `xlnet` can be applied either unidirectionally or bidirectionally.

# Sentence Probabilities for Transformers

- `lstm` and `gpt2` are unidirectional, and so they can be used to compute the probability of a sentence left to right, according to the formula $\overrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i})$.
- `bert` is bidirectional, and predicts words for both their left and right contexts.
- It requires the formula $\overleftrightarrow{P}(s) = \prod_{i=0}^{|s|} P(w_i | w_{<i}, w_{>i})$.
- This equation does not yield true probabilities, as its values cannot be normalised to sum to 1.
- Instead these values provide confidence scores of likelihood.
- `xlnet` can be applied either unidirectionally or bidirectionally.

# Language Model Architectures

| Model | Configuration | | | Training Data | | | |
|---|---|---|---|---|---|---|---|
| | Architecture | Encoding | #Param. | Casing | Size | Tokenisation | Corpora |
| lstm | RNN | Unidir. | 60M | Uncased | 0.2GB | Word | Wikipedia |
| tdlm | RNN | Unidir. | 80M | Uncased | 0.2GB | Word | Wikipedia |
| gpt2 | Transformer | Unidir. | 340M | Cased | 40GB | BPE | WebText |
| $\texttt{bert}_{\texttt{cs}}$ | Transformer | Bidir. | 340M | Cased | 13GB | WordPiece | Wikipedia, BookCorpus |
| $\texttt{bert}_{\texttt{ucs}}$ | Transformer | Bidir. | 340M | Uncased | 13GB | WordPiece | Wikipedia, BookCorpus |
| xlnet | Transformer | Hybrid | 340M | Cased | 126GB | Sentence-Piece | Wikipedia, BookCorpus, Giga5 ClueWeb, Common Crawl |

# Acceptability Scoring Measures

| Acc. Measure | Equation |
|---|---|
| *LogProb* | $\log P_m(s)$ |
| *Mean LP* | $\frac{\log P_m(s)}{\lvert s\rvert}$ |
| *PenLP* | $\frac{\log P_m(s)}{((5+\lvert s\rvert)/(5+1))^{\alpha}}$ |
| *NormLP* | $-\frac{\log P_m(s)}{\log P_u(s)}$ |
| *SLOR* | $\frac{\log P_m(s)-\log P_u(s)}{\lvert s\rvert}$ |

$P(s)$ is the sentence probability, computed using either the uni-prob or bi-prob formula, depending on the model, $P_u(s)$ is the sentence probability estimated by a unigram language model, and $\alpha = 0.8$.

# Upper Bounds on Model Performance

- We compute two human performance estimates to serve as upper bounds on the accuracy of a model.
- $\mathrm{ub}_1$ is the one-vs-rest annotator correlation, where we select a random annotator's rating and compare it to the mean rating of the rest, using Pearson's $r$.
- We repeat this for a large number of trials (1000) to get a robust estimate of the mean correlation.
- $\mathrm{ub}_2$ is the half-vs-half annotator correlation, where for each sentence we randomly split the annotators into two groups, and compare the mean ratings between the groups.
- The simulated human performance is fairly consistent over context types, with $\mathrm{ub}_1 = 0.66$, 0.65, and 0.68 for $\mathrm{h}^{\varnothing}$, $\mathrm{h}^{+}$, and $\mathrm{h}^{-}$, respectively.

# Upper Bounds on Model Performance

- We compute two human performance estimates to serve as upper bounds on the accuracy of a model.
- $\mathtt{ub}_1$ is the one-vs-rest annotator correlation, where we select a random annotator's rating and compare it to the mean rating of the rest, using Pearson's *r*.
- We repeat this for a large number of trials (1000) to get a robust estimate of the mean correlation.
- $\mathtt{ub}_2$ is the half-vs-half annotator correlation, where for each sentence we randomly split the annotators into two groups, and compare the mean ratings between the groups.
- The simulated human performance is fairly consistent over context types, with $\mathtt{ub}_1 = 0.66$, 0.65, and 0.68 for $\mathtt{h}^{\varnothing}$, $\mathtt{h}^{+}$, and $\mathtt{h}^{-}$, respectively.

# Upper Bounds on Model Performance

- We compute two human performance estimates to serve as upper bounds on the accuracy of a model.
- $\mathtt{ub}_1$ is the one-vs-rest annotator correlation, where we select a random annotator's rating and compare it to the mean rating of the rest, using Pearson's *r*.
- We repeat this for a large number of trials (1000) to get a robust estimate of the mean correlation.
- $\mathtt{ub}_2$ is the half-vs-half annotator correlation, where for each sentence we randomly split the annotators into two groups, and compare the mean ratings between the groups.
- The simulated human performance is fairly consistent over context types, with $\mathtt{ub}_1 = 0.66$, 0.65, and 0.68 for $\mathtt{h}^{\varnothing}$, $\mathtt{h}^{+}$, and $\mathtt{h}^{-}$, respectively.

# Upper Bounds on Model Performance

- We compute two human performance estimates to serve as upper bounds on the accuracy of a model.
- $\mathtt{ub}_1$ is the one-vs-rest annotator correlation, where we select a random annotator's rating and compare it to the mean rating of the rest, using Pearson's *r*.
- We repeat this for a large number of trials (1000) to get a robust estimate of the mean correlation.
- $\mathtt{ub}_2$ is the half-vs-half annotator correlation, where for each sentence we randomly split the annotators into two groups, and compare the mean ratings between the groups.
- The simulated human performance is fairly consistent over context types, with $\mathtt{ub}_1 = 0.66$, 0.65, and 0.68 for $\mathtt{h}^{\varnothing}$, $\mathtt{h}^{+}$, and $\mathtt{h}^{-}$, respectively.

# Upper Bounds on Model Performance

- We compute two human performance estimates to serve as upper bounds on the accuracy of a model.
- $\mathtt{ub}_1$ is the one-vs-rest annotator correlation, where we select a random annotator's rating and compare it to the mean rating of the rest, using Pearson's *r*.
- We repeat this for a large number of trials (1000) to get a robust estimate of the mean correlation.
- $\mathtt{ub}_2$ is the half-vs-half annotator correlation, where for each sentence we randomly split the annotators into two groups, and compare the mean ratings between the groups.
- The simulated human performance is fairly consistent over context types, with $\mathtt{ub}_1 = 0.66$, 0.65, and 0.68 for $\mathtt{h}^{\varnothing}$, $\mathtt{h}^{+}$, and $\mathtt{h}^{-}$, respectively.

# Model Performance: Null Context

| Rtg | Encod. | Model | *LogProb* | *Mean LP* | *PenLP* | *NormLP* | *SLOR* |
|---|---|---|---|---|---|---|---|
| $h^{\varnothing}$ | Unidir. | $\mathtt{lstm}^{\varnothing}$ | 0.28 | 0.42 | 0.42 | 0.53 | **0.54** |
| | | $\mathtt{lstm}^{+}$ | 0.30 | 0.50 | 0.45 | 0.62 | **0.64** |
| | | $\mathtt{tdlm}^{\varnothing}$ | 0.29 | 0.50 | 0.45 | 0.61 | **0.62** |
| | | $\mathtt{tdlm}^{+}$ | 0.30 | 0.51 | 0.46 | 0.61 | **0.62** |
| | | $\mathtt{gpt2}^{\varnothing}$ | 0.33 | 0.43 | **0.55** | 0.51 | 0.51 |
| | | $\mathtt{gpt2}^{+}$ | 0.38 | 0.58 | 0.60 | **0.63** | 0.62 |
| | | $\mathtt{xlnet}^{\varnothing}_{\mathtt{uni}}$ | 0.30 | 0.43 | 0.51 | 0.52 | **0.53** |
| | | $\mathtt{xlnet}^{+}_{\mathtt{uni}}$ | 0.36 | 0.58 | 0.56 | 0.62 | **0.63** |
| | Bidir. | $\mathtt{bert}^{\varnothing}_{\mathtt{cs}}$ | 0.50 | 0.55 | **0.63** | 0.56 | 0.53 |
| | | $\mathtt{bert}^{+}_{\mathtt{cs}}$ | 0.53 | 0.63 | **0.67** | 0.64 | 0.60 |
| | | $\mathtt{bert}^{\varnothing}_{\mathtt{ucs}}$ | 0.59 | 0.63 | **0.70** | 0.63 | 0.60 |
| | | $\mathtt{bert}^{+}_{\mathtt{ucs}}$ | 0.60 | 0.68 | **0.73** | 0.68 | 0.64 |
| | | $\mathtt{xlnet}^{\varnothing}_{\mathtt{bi}}$ | 0.52 | 0.52 | **0.66** | 0.53 | 0.54 |
| | | $\mathtt{xlnet}^{+}_{\mathtt{bi}}$ | 0.58 | 0.66 | **0.74** | 0.67 | 0.66 |
| | — | $\mathtt{ub}_1$ | | | 0.66 | | |
| | | $\mathtt{ub}_2$ | | | 0.88 | | |

# Model Performance: Real Context

| Rtg | Encod. | Model | *LogProb* | *Mean LP* | *PenLP* | *NormLP* | *SLOR* |
|---|---|---|---|---|---|---|---|
| $h^+$ | Unidir. | $\mathtt{lstm}^{\varnothing}$ | 0.29 | 0.44 | 0.43 | 0.52 | **0.53** |
| | | $\mathtt{lstm}^{+}$ | 0.31 | 0.52 | 0.46 | **0.63** | **0.63** |
| | | $\mathtt{tdlm}^{\varnothing}$ | 0.29 | 0.50 | 0.45 | **0.60** | 0.59 |
| | | $\mathtt{tdlm}^{+}$ | 0.30 | 0.51 | 0.46 | **0.59** | **0.59** |
| | | $\mathtt{gpt2}^{\varnothing}$ | 0.33 | 0.43 | **0.55** | 0.50 | 0.50 |
| | | $\mathtt{gpt2}^{+}$ | 0.38 | 0.59 | 0.61 | **0.63** | 0.61 |
| | | $\mathtt{xlnet}^{\varnothing}_{\mathtt{uni}}$ | 0.30 | 0.43 | **0.51** | 0.50 | **0.51** |
| | | $\mathtt{xlnet}^{+}_{\mathtt{uni}}$ | 0.36 | 0.57 | 0.56 | **0.61** | **0.61** |
| | Bidir. | $\mathtt{bert}^{\varnothing}_{\mathtt{cs}}$ | 0.49 | 0.54 | **0.62** | 0.54 | 0.51 |
| | | $\mathtt{bert}^{+}_{\mathtt{cs}}$ | 0.52 | 0.63 | **0.67** | 0.63 | 0.58 |
| | | $\mathtt{bert}^{\varnothing}_{\mathtt{ucs}}$ | 0.58 | 0.63 | **0.70** | 0.63 | 0.59 |
| | | $\mathtt{bert}^{+}_{\mathtt{ucs}}$ | 0.60 | 0.68 | **0.73** | 0.67 | 0.63 |
| | | $\mathtt{xlnet}^{\varnothing}_{\mathtt{bi}}$ | 0.51 | 0.51 | **0.66** | 0.52 | 0.52 |
| | | $\mathtt{xlnet}^{+}_{\mathtt{bi}}$ | 0.58 | 0.65 | **0.74** | 0.66 | 0.65 |
| | — | $\mathtt{ub}_1$ | | | 0.65 | | |
| | | $\mathtt{ub}_2$ | | | 0.89 | | |

# Model Performance: Random Context

| Rtg | Encod. | Model | *LogProb* | *Mean LP* | *PenLP* | *NormLP* | *SLOR* |
|---|---|---|---|---|---|---|---|
| $h^-$ | Unidir. | $\mathtt{lstm}^{\varnothing}$ | 0.29 | 0.45 | 0.44 | **0.51** | 0.50 |
| | | $\mathtt{lstm}^{-}$ | 0.28 | 0.41 | 0.41 | **0.48** | **0.48** |
| | | $\mathtt{tdlm}^{\varnothing}$ | 0.30 | 0.52 | 0.46 | **0.60** | 0.58 |
| | | $\mathtt{tdlm}^{-}$ | 0.29 | 0.49 | 0.45 | **0.56** | **0.56** |
| | | $\mathtt{gpt2}^{\varnothing}$ | 0.33 | 0.43 | **0.55** | 0.49 | 0.47 |
| | | $\mathtt{gpt2}^{-}$ | 0.31 | 0.40 | **0.52** | 0.45 | 0.44 |
| | | $\mathtt{xlnet}_{\mathtt{uni}}^{\varnothing}$ | 0.31 | 0.44 | **0.52** | 0.49 | 0.50 |
| | | $\mathtt{xlnet}_{\mathtt{uni}}^{-}$ | 0.30 | 0.41 | **0.50** | 0.47 | 0.47 |
| | Bidir. | $\mathtt{bert}_{\mathtt{cs}}^{\varnothing}$ | 0.49 | 0.53 | **0.62** | 0.53 | 0.49 |
| | | $\mathtt{bert}_{\mathtt{cs}}^{-}$ | 0.50 | 0.52 | **0.61** | 0.51 | 0.47 |
| | | $\mathtt{bert}_{\mathtt{ucs}}^{\varnothing}$ | 0.57 | 0.61 | **0.69** | 0.60 | 0.56 |
| | | $\mathtt{bert}_{\mathtt{ucs}}^{-}$ | 0.56 | 0.58 | **0.67** | 0.57 | 0.54 |
| | | $\mathtt{xlnet}_{\mathtt{bi}}^{\varnothing}$ | 0.50 | 0.48 | **0.63** | 0.49 | 0.49 |
| | | $\mathtt{xlnet}_{\mathtt{bi}}^{-}$ | 0.51 | 0.51 | **0.65** | 0.52 | 0.51 |
| | — | $\mathtt{ub}_1$ | | | 0.68 | | |
| | | $\mathtt{ub}_2$ | | | 0.88 | | |

# Modelling Experiment Results

- The bidirectional models significantly outperform the unidirectional models across all three context types, when *PenLP*, rather than *SLOR* is the scoring function.
- This suggests that large lexical embeddings and bidirectional context training render normalisation by word frequency unnecessary.
- Model architecture rather than size is the decisive factor governing performance.
- $\texttt{bert}_{\texttt{ucs}}$ surpasses estimated individual human performance, as specified by $\texttt{ub}_1$, on the the prediction of sentence acceptability task.

# Modelling Experiment Results

- The bidirectional models significantly outperform the unidirectional models across all three context types, when *PenLP*, rather than *SLOR* is the scoring function.
- This suggests that large lexical embeddings and bidirectional context training render normalisation by word frequency unnecessary.
- Model architecture rather than size is the decisive factor governing performance.
- $\texttt{bert}_{\texttt{ucs}}$ surpasses estimated individual human performance, as specified by $\texttt{ub}_1$, on the the prediction of sentence acceptability task.

# Modelling Experiment Results

- The bidirectional models significantly outperform the unidirectional models across all three context types, when *PenLP*, rather than *SLOR* is the scoring function.
- This suggests that large lexical embeddings and bidirectional context training render normalisation by word frequency unnecessary.
- Model architecture rather than size is the decisive factor governing performance.
- $\texttt{bert}_{\texttt{ucs}}$ surpasses estimated individual human performance, as specified by $\texttt{ub}_1$, on the the prediction of sentence acceptability task.

# Modelling Experiment Results

- The bidirectional models significantly outperform the unidirectional models across all three context types, when *PenLP*, rather than *SLOR* is the scoring function.
- This suggests that large lexical embeddings and bidirectional context training render normalisation by word frequency unnecessary.
- Model architecture rather than size is the decisive factor governing performance.
- $\texttt{bert}_{\texttt{ucs}}$ surpasses estimated individual human performance, as specified by $\texttt{ub}_1$, on the the prediction of sentence acceptability task.

# Conclusions

- Processing context induces a cognitive load for humans, which creates a compression effect on the distribution of acceptability ratings.
- If the context is relevant to the sentence, a discourse coherence effect uniformly boosts sentence acceptability.
- Bidirectional neural language models outperform unidirectional models on the sentence acceptability prediction task.
- Our best bidirectional model surpasses estimated individual human performance on this task.

# Conclusions

- Processing context induces a cognitive load for humans, which creates a compression effect on the distribution of acceptability ratings.
- If the context is relevant to the sentence, a discourse coherence effect uniformly boosts sentence acceptability.
- Bidirectional neural language models outperform unidirectional models on the sentence acceptability prediction task.
- Our best bidirectional model surpasses estimated individual human performance on this task.

# Conclusions

- Processing context induces a cognitive load for humans, which creates a compression effect on the distribution of acceptability ratings.
- If the context is relevant to the sentence, a discourse coherence effect uniformly boosts sentence acceptability.
- Bidirectional neural language models outperform unidirectional models on the sentence acceptability prediction task.
- Our best bidirectional model surpasses estimated individual human performance on this task.

# Conclusions

- Processing context induces a cognitive load for humans, which creates a compression effect on the distribution of acceptability ratings.
- If the context is relevant to the sentence, a discourse coherence effect uniformly boosts sentence acceptability.
- Bidirectional neural language models outperform unidirectional models on the sentence acceptability prediction task.
- Our best bidirectional model surpasses estimated individual human performance on this task.

# Future Work

- We will consider alternative ways to present sentences for acceptability judgments.
- We plan to extend `tdlm` by incorporating a bidirectional design, as this architecture has been shown to be promising.
- We will extend our experiments to other languages.
- We intend to explore the impact of different sorts of contexts, both linguistic and non-linguistic, on this task.

# Future Work

- We will consider alternative ways to present sentences for acceptability judgments.
- We plan to extend `tdlm` by incorporating a bidirectional design, as this architecture has been shown to be promising.
- We will extend our experiments to other languages.
- We intend to explore the impact of different sorts of contexts, both linguistic and non-linguistic, on this task.

# Future Work

- We will consider alternative ways to present sentences for acceptability judgments.
- We plan to extend `tdlm` by incorporating a bidirectional design, as this architecture has been shown to be promising.
- We will extend our experiments to other languages.
- We intend to explore the impact of different sorts of contexts, both linguistic and non-linguistic, on this task.

# Future Work

- We will consider alternative ways to present sentences for acceptability judgments.
- We plan to extend `tdlm` by incorporating a bidirectional design, as this architecture has been shown to be promising.
- We will extend our experiments to other languages.
- We intend to explore the impact of different sorts of contexts, both linguistic and non-linguistic, on this task.